Solution 1: Bayesian Linear Model

The posterior distribution is obtained by Bayes' rule

$$\underbrace{p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{q(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y}|\mathbf{X})}_{\text{marginal}}}.$$

In the Bayesian linear model we have a Gaussian likelihood: $\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I}_n)$, i.e.,

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) &\propto \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})\right] \\ &= \exp\left[-\frac{\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2}{2\sigma^2}\right] \\ &= \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right]. \end{aligned}$$

Moreover, note that the maximum a posteriori estimate of $\boldsymbol{\theta}$, which is defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})$$

can also be defined by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\right),$$

since log is a monotonically increasing function, so the maximizer is the same.

(a) If the prior distribution is a uniform distribution over the parameter vectors $\boldsymbol{\theta}$, i.e.,

$$q(\boldsymbol{\theta}) \propto 1,$$

then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) q(\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2}\right]. \end{aligned}$$

With this,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{y})\right) \\ &= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\ &= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2}{2\sigma^2} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2, \qquad (2\sigma^2 \text{ is just a constant scaling}) \end{aligned}$$

so the maximum a posteriori estimate coincides with the empirical risk minimizer for the L2-loss (over the linear models).

(b) If we choose a Gaussian distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right], \qquad \tau > 0,$$

then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{1}{2\tau^2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right] \\ &= \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2}\right] \end{aligned}$$

With this,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})\right) \\ &= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2} \\ &= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{\|\boldsymbol{\theta}\|_2^2}{2\tau^2} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2 + \frac{\sigma^2}{\tau^2}\|\boldsymbol{\theta}\|_2^2, \end{aligned}$$

so the maximum a posteriori estimate coincides for the choice of $\lambda = \frac{\sigma^2}{\tau^2} > 0$ with the regularized empirical risk minimizer for the L2-loss with L2 penalty (over the linear models), i.e., the Ridge regression.

(c) If we choose a Laplace distribution over the parameter vectors $\boldsymbol{\theta}$ as the prior belief, i.e.,

$$q(\boldsymbol{\theta}) \propto \exp\left[-\frac{\sum_{i=1}^p |\boldsymbol{\theta}_i|}{\tau}\right], \qquad \tau > 0,$$

then

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{X},\boldsymbol{\theta})q(\boldsymbol{\theta}) \\ &\propto \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\sum_{i=1}^p |\boldsymbol{\theta}_i|}{\tau}\right] \\ &= \exp\left[-\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_1}{\tau}\right] \end{aligned}$$

With this,

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \arg\max_{\boldsymbol{\theta}} \log\left(p(\boldsymbol{\theta}|\mathbf{X},\mathbf{y})\right) \\ &= \arg\max_{\boldsymbol{\theta}} -\frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} - \frac{\|\boldsymbol{\theta}\|_1}{\tau} \\ &= \arg\min_{\boldsymbol{\theta}} \frac{\sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2}{2\sigma^2} + \frac{\|\boldsymbol{\theta}\|_1}{\tau} \\ &= \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^n (y^{(i)} - \boldsymbol{\theta}^\top\mathbf{x}^{(i)})^2 + \frac{2\sigma^2}{\tau}\|\boldsymbol{\theta}\|_1, \end{aligned}$$

so the maximum a posteriori estimate coincides for the specific choice of $\lambda = \frac{2\sigma^2}{\tau}$ with the regularized empirical risk minimizer for the L2-loss with L1 penalty (over the linear models), i.e., the Lasso regression.

Exercise 1: Covariance Functions

(a) Proofs:

(i) $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is a valid covariance function since the kernel matrix $\mathbf{K} = \sigma_0^2 \cdot \mathbf{1}\mathbf{1}^T$ is a positive semi-definite matrix. This can be proved as follows: first, $\mathbf{K} = \mathbf{K}^T$ is symmetric; second, for all $\mathbf{v} = (v_1, \ldots, v_n)^T \in \mathbb{R}^n$, $\mathbf{v}^T\mathbf{K}\mathbf{v} = \sigma_0^2 \cdot (\sum_{i=1}^n v_i, \ldots, \sum_{i=1}^n v_i) \cdot (v_1, \ldots, v_n)^T = \sigma_0^2 \cdot [v_1 \sum_{i=1}^n v_i + \ldots + v_n \sum_{i=1}^n v_n] = \sigma_0^2 \cdot [(\sum_{i=1}^n v_i) \cdot (\sum_{i=1}^n v_i)] \geq 0$.

(ii) To prove that $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^T\mathbf{x}'$ is a valid covariance function, we need to notice that σ_0^2 and $\mathbf{x}^T\mathbf{x}'$ are both valid covariance function, and sum operation also yield a valid covariance function.

(iii) $k(\mathbf{x}, \mathbf{x}') = (\sigma_0^2 + \mathbf{x}^T\mathbf{x}')^p$ is a valid covariance function since the linear function is a covariance function, and the only polynomial coefficient 1 is positive.

(iv) The squared exponential can be written as $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\ell^2}) \cdot \exp(\frac{\mathbf{x}^T\mathbf{x}'}{\ell^2}) \cdot \exp(-\frac{\mathbf{x}'^T\mathbf{x}'}{2\ell^2})$. Note that $\exp(\frac{\mathbf{x}^T\mathbf{x}'}{\ell^2})$ is a valid covariance function and can be easily proved using the composition rules. We further define $t(\mathbf{x}) = \exp(-\frac{\mathbf{x}^T\mathbf{x}}{2\ell^2})$. Therefore, $k(\mathbf{x}, \mathbf{x}') = t(\mathbf{x}) \cdot \exp(\frac{\mathbf{x}^T\mathbf{x}'}{\ell^2}) \cdot t(\mathbf{x}')$ is a valid covariance function.

(b) $k(\cdot, \cdot)$ is called stationary if $k(\mathbf{x}, \mathbf{x} + \mathbf{d}) = k(\mathbf{0}, \mathbf{d})$; $k(\cdot, \cdot)$ is called isotropic if it is a function of $||\mathbf{x} - \mathbf{x}'||$.

(i) $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2$ is stationary since $k(\mathbf{x}, \mathbf{x}+\mathbf{d}) = k(\mathbf{0}, \mathbf{d}) = \sigma_0^2$. It can be written as $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 ||\mathbf{x} - \mathbf{x}'||^0$, so it is isotropic.

(ii) $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x}^T\mathbf{x}'$ is not stationary since $k(\mathbf{x}, \mathbf{x} + \mathbf{d}) = \sigma_0^2 + \mathbf{x}^T\mathbf{x} + \mathbf{x}^T\mathbf{d}$, while $k(\mathbf{0}, \mathbf{x} + \mathbf{d}) = \sigma_0^2$. Furthermore, it can not be written as $k(||\mathbf{x} - \mathbf{x}'||)$, so it is not isotropic.

(iii) Similar to linear covariance function, the polynomial covariance function is neither stationary nor isotropic.

(iv) The squared exponential covariance function is stationary as $k(\mathbf{x}, \mathbf{x} + \mathbf{d}) = k(\mathbf{0}, \mathbf{d}) = \exp(-\frac{||d||^2}{2\ell^2})$. It is a function of $||\mathbf{x} - \mathbf{x}'||$, so it is isotropic.

(v) Similar to the argument of squared exponential covariance function, the Matérn covariance function is stationary and isotropic.

(vi) Similar to the argument of squared exponential covariance function, the exponential covariance function is stationary and isotropic.